PHIL 7001: Fundamentals of AI, Data, and Algorithms

Week 6
Linear (Regression) Models

Boris Babic,
HKU 100 Associate Professor of Data Science, Law and Philosophy

Today

Learning goals

- Understand the basic concepts of linear models.
- Learn how to interpret and analyze the output of a linear regression model.
- Understand linear model functions in R
- Familiarize yourself with the `lm()` function in R for creating linear models.
- Gain insight into accessing and interpreting estimated coefficients using the `summary()` function.
- Apply your knowledge to a practical exercise to predict outcomes using linear regression models.

Last week

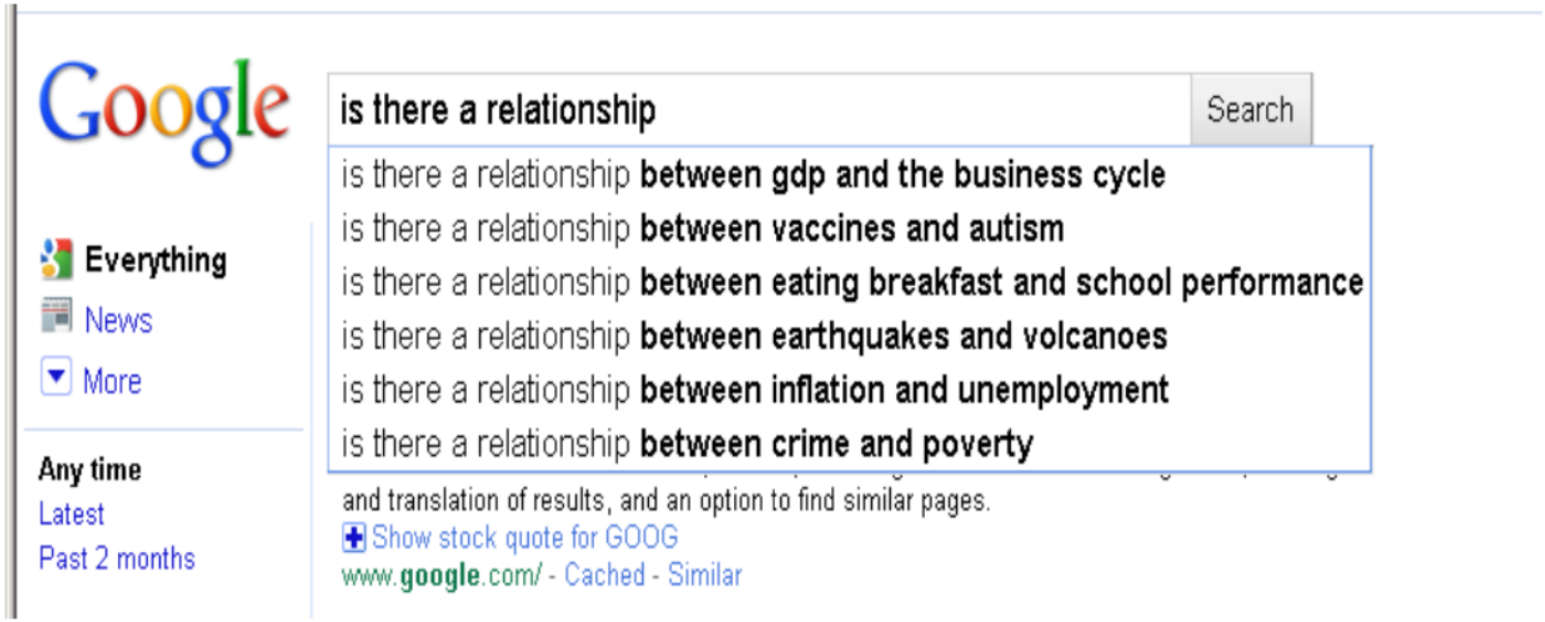
Review of last week

- Introduction to estimation
- Overview of hypothesis tests
- P-values
- Bayesian statistics (will not be tested)

DO YOU BELIEVE THAT...

- Children watching violent movies become more violent
- Legalizing abortion reduces crime
- People live longer if they eat more brocoli /sleep less
- Being single is worse for you than smoking (men vs. women?)

RELATIONSHIPS: WHAT DO PEOPLE WANT TO KNOW



CORRELATION – WEATHER AND STOCK RETURNS

THE JOURNAL OF FINANCE • VOL. LVIII, NO. 3 • JUNE 2003

Good Day Sunshine: Stock Returns and the Weather

DAVID HIRSHLEIFER and TYLER SHUMWAY*

ABSTRACT

Psychological evidence and casual intuition predict that sunny weather is associated with upbeat mood. This paper examines the relationship between morning sunshine in the city of a country's leading stock exchange and daily market index returns across 26 countries from 1982 to 1997. Sunshine is strongly significantly correlated with stock returns. After controlling for sunshine, rain and snow are unrelated to returns. Substantial use of weather-based strategies was optimal for a trader with very low transactions costs. However, because these strategies involve frequent trades, fairly modest costs eliminate the gains. These findings are difficult to reconcile with fully rational price setting.

CORRELATION – MOOD AND STOCK RETURNS

Football and stock returns*

Alex Edmans
Sloan School of Management at MIT

Diego García
Tuck School of Business at Dartmouth

Øyvind Norli
Tuck School of Business at Dartmouth

Abstract

This paper investigates the stock market reaction to the outcome of international football competitions, such as the FIFA World Cup, a variable shown in psychological literature to have a dramatic effect on mood. We document an economically and statistically significant market decline after football losses. Daily stock returns are 39 basis points lower than average following a loss in a World Cup elimination match. This football-loss effect is robust to changes in estimation methodology and to the removal of outliers in the data. It is particularly strong for more critical games, in recent years, and in countries where football is especially important. Controlling for the pre-game expected outcome, we are able to reject that the football-loss effect is caused by economic factors such as reduced productivity or lost revenues. Coupled with the size of the effect and its concentration in small stocks, we suggest that the football-loss effect stems from the impact of football on investor mood.

Introduction to Regression Analysis

Regression analysis is used to:

- **Predict the value of a dependent variable based on at least one independent variable**
- **Explain the impact of changes in an independent variable on the dependent variable**

Jargon:

- **Dependent variable: The variable we wish to predict or explain**
- **Independent variable: The variable used to explain the dependent variable**

Types of regression models:

- **Simple Regression: Use one independent variable to predict another (today)**
- **Multiple Regression: Use more than one independent variable to predict another variable**

- Draw a line through these points that capture the relationship between two variables.

Whether our goal is to use Height to (1) explain the variation in weight or (2) predict Weight, we need to build a model.

One option is a simple linear regression model, which assumes that there is a "best" straight line that explains the true or real relationship between X (predictor/dependent variable) and y (response/independent variable), and that the values we observe randomly deviate from this line.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Y_i : response variable (or dependent variable, target variable, ...) for the i^{th} observation
- x_i : independent variable (or predictor, covariate, feature, input, ...) for the i^{th} observation
- β_0 : intercept parameter (where the line crosses through the y-axis.
- β_1 : slope parameter. This is the coefficient associated with the variable x. If it is larger in magnitude, then x has a stronger effect on the response.
- ϵ_i : random error term for i^{th} observation.

Imagine we had data on the entire population...

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- There is a true line with slope β_1 and intercept β_0 which describes the overall relationship between x and y.
- There are random deviations between this line and the particular individual observations y_i.

True vs. Fitted

Postulated Model vs. Estimated Model

Postulated Model (for the *population*)

$$Y_i = A + BX_i + e_i$$

Definitions:

- Y Dependent variable
- X Independent variable
- A, B Unknown Regression Parameters
- e_i Random Error Term

Assumptions:

- $e_i \sim Normal\,(0, \sigma^2)$
- e_i's are independent

Goal: Estimate A and B (based on a sample)

Estimated Model (based on a *sample*)

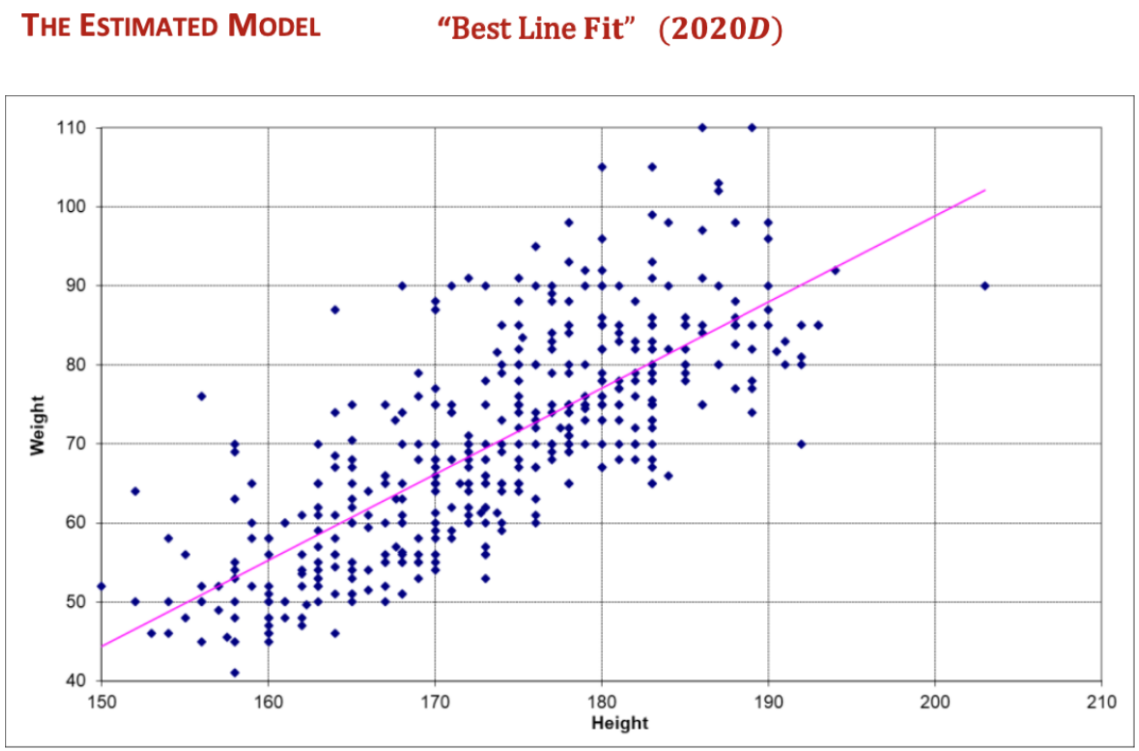
$$\widehat{Y_i} = a + bX_i$$

Definitions:

- a, b Regression Coefficients (estimates of Regression Parameters)
- $\widehat{Y}$ is the Predicted Value of Y for a given X

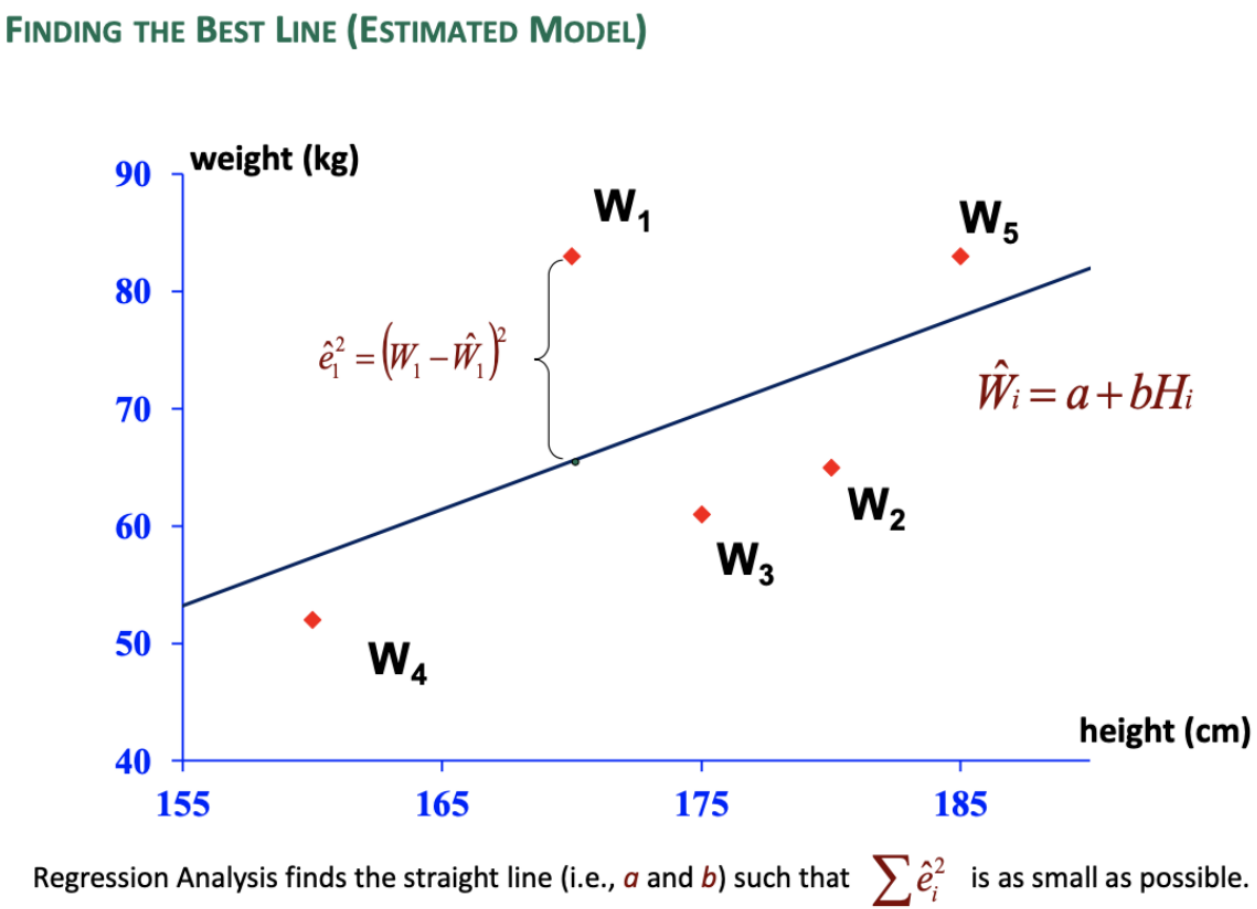
Simple linear regression: What is the best line?

- When we have a data from a sample, we want to find a line which is as close as possible to as many points as possible.
- What do we mean by "as close as possible"?



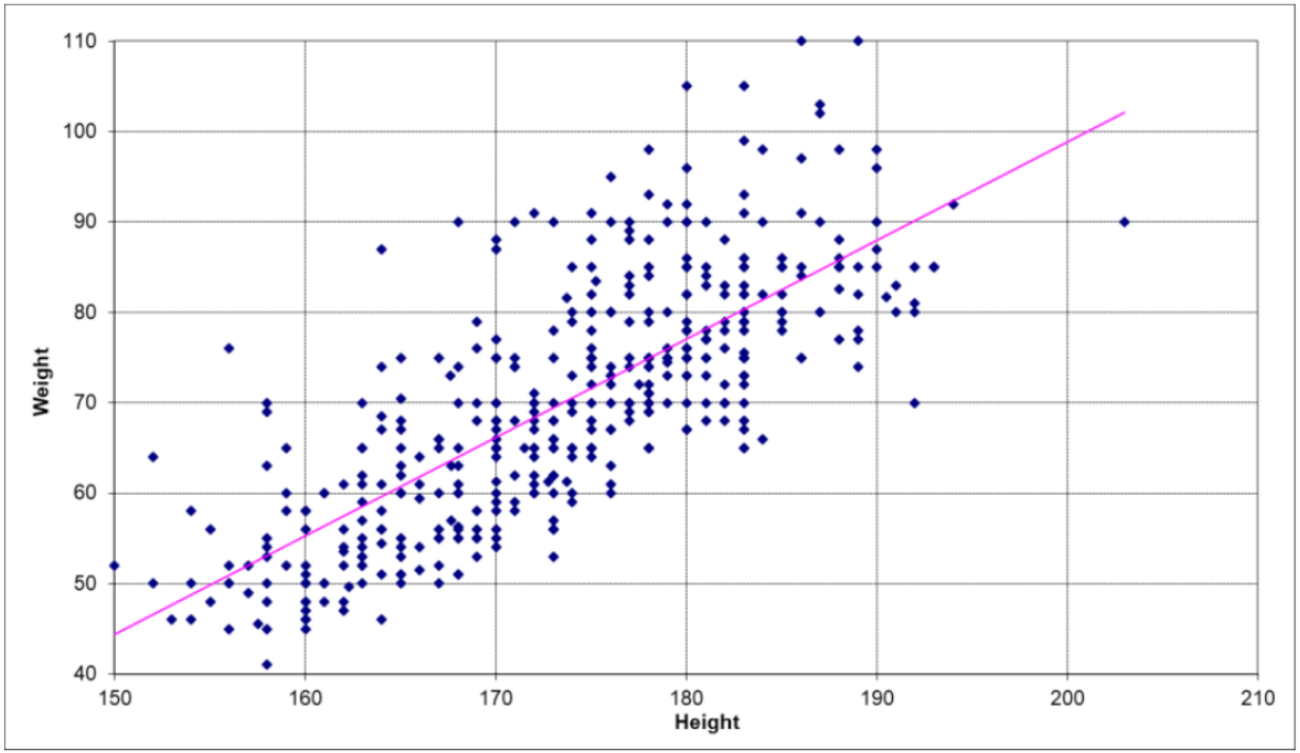
Least squares Regression Line

- Sum of vertical distances between the points and the line (i.e. squared dierences between height of point and line) is minimized



THE ESTIMATED MODEL

$$\widehat{Weight}_i = -118.93 + 1.09(Height_i)$$



Interpreting the slope and intercept with a numerical predictor

The estimated simple linear regression of weight on height (i.e., the fitted line) is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{y}$ (called the fitted or predicted value) is the estimated average value of y when the predictor is equal to x.

- The slope $\hat{\beta}_1$ is the average change in y for a 1-unit change in x
- The intercept $\hat{\beta}_0$ is the average of y when $x_i = 0$ (often this doesn't make sense, but tells us the height of the line).
- The difference between the observed and predicted value of y for the i^{th} observation is called the residual $e_i = y_i - \hat{y}_i$

Interpreting the slope and intercept with a numerical predictor

Weight versus Height

- **Estimates a linear relationship which corresponds to a straight line that goes through the data**

$$\widehat{Weight}_i = a + b(Height_i)$$

$$\widehat{Weight}_i = -118.93 + 1.09(Height_i)$$

Dependent variable

intercept (constant)

coefficient (slope)

Independent variable (Explanatory variable)

What does this mean?

Each *cm* of Height increases Weight by 1.05 *kg*, on average.

Regression: Are the Predictors Significant?

- ***Question:*** ***Does the independent (predictor X) variable have an impact on the dependent variable Y? In other words, is the true slope coefficient really different from zero ?***

  In general, the estimated coefficient will almost surely not be *exactly* zero. Is this due to the just chance (randomness), or is there really a relationship?

- ***Measure*****: p-value of the coefficient tells us how significant the effect is, based on the data**

  The lower the p-value, the more likely it is that the slope is really different from zero, i.e. that there is a significant impact of the explanatory variable.

- ***Test*** **(at α = 5% significance level):**

  If p-value < 5%, then there is a significant impact!

- ***Action*****: If the predictor variable *X* is not significant, then remove it from the analysis!**

Example

- Recall that in Lecture 1 we had a scenario that studied the relationship between the amount of exercise people get in a week and their body fat percentage.
- We have collected data from a group of men and women who have different levels of exercise each week. Our goal is to determine if there's a correlation between the number of hours spent exercising and body fat percentage in these two groups.

```
Rows: 25
Columns: 3
$ exercise <dbl> 6, 2, 7, 8, 5, 7, 8, 6, 7, 3, 6, 6, 1, 2, 3, 4, 2, 3, 4, 2, 1, 1, 3, 5, 6
$ body_fat <dbl> 20, 32, 15, 10, 25, 15, 12, 22, 18, 23, 13, 15, 29, 30, 25, 28, 29, 29, 18, 26, 25, 27, 24, 15, 18
$ gender   <chr> "Men", "Men", "Men", "Men", "Men", "Men", "Men", "Men", "Men", "Men", "Men", "Men", "Women", "Women", "Women"
```

Example

Let's build a simple regression model to investigate the relationship between the number of hours spent exercising and body fat percentage.

- In R, this can be done with the **'lm()'** function
- The **'lm()'** function forms the core of linear modeling in R.
- Syntax: lm(formula, data)
- Parameters
  - formula: Specifies the relationship between variables using a formula.
  - data: The dataset containing the variables.

Example

In our scenario,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The dependent variable (y) will be the body fat percentage.
- The independent variable (x), also called the predictor, will be the number of hours spent exercising.
- In R, we can build the model by calling the **lm()** function

```
# Fit the model for all people
# The response variable is the body fat percentage
# while the predictors are exercise hours
model_all <- lm(body_fat ~ exercise, data=my_data)
```

Example

- In R, the **summary()** function provides a comprehensive summary of the linear model.
- This summary includes coefficients, standard errors, t-values, p-values, residuals, and R-squared.

```r
# Print the summary of the model
summary(model_all)
```

```
Call:
lm(formula = body_fat ~ exercise, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0733 -2.6518 -0.2302  2.6620  5.5051

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.1811     1.4480  22.224  < 2e-16 ***
exercise     -2.4216     0.2976  -8.138  3.2e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.333 on 23 degrees of freedom
Multiple R-squared:  0.7422,    Adjusted R-squared:  0.731
F-statistic: 66.22 on 1 and 23 DF,  p-value: 3.196e-08
```

Example

To specifically access the estimated coefficients, you can use **summary()$coefficients**.

```
# Only print the coefficienta summary of the model
summary(model_all)$coefficients
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 32.181122  1.4480290 22.224087 4.797699e-17
exercise    -2.421556  0.2975681 -8.137821 3.195662e-08
```

Example - Model Summary Output

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 32.181122  1.4480290 22.224087 4.797699e-17
exercise    -2.421556  0.2975681 -8.137821 3.195662e-08
```

- **(Intercept)** is the estimate of β_0 (i.e. $\hat{\beta}_0$), which is 32.181122.
- **exercise** is the estimate of β_1 (i.e. $\hat{\beta}_1$), which is -2.421556.
- **Std. Error** are standard errors that indicate the variability of the estimates.
- **t value and** $Pr(>|t|)$ measure the significance of estimates. $Pr(>|t|)$ is the p-values. Lower p-values suggest more significant relationships.

Example - Interpreting Model Summary

Gaining Profound Insights from the Model Summary

- **Coefficient $\hat{\beta}_0$:** The intercept (32.181122) is the baseline body fat percentage when exercise hours are zero.
- **Coefficient $\hat{\beta}_1$:** For each additional hour of exercise, body fat percentage decreases by 2.421556 units.
- **Standard Error:** The standard errors (1.4480290 for intercept, 0.2975681 for exercise) indicate the variability of the estimates.
- $Pr(>|t|)$**:** The p-values is 3.195662e-08 for exercise, which is much lower than 0.05, indicating the significance of the relationship between exercise hours and body fat percentage.

Example - Predictions

- We are also able to predict the body fat percentage given one's exercise level.
- Questions: Now, if we know that a student spends 5 hours per week on exercise, what would be his predicted body fat percentage?

Example - Predictions

- We are also able to predict the body fat percentage given one's exercise level.
- Questions: Now, if we know that a student spends 5 hours per week on exercise, what would be his predicted body fat percentage?

$$\hat{y} = 32.181122 - 2.421556 * 5 \approx 20.073$$

- The student has a body fat percentage of 20.073%

Example: Model Visualization

Now, let's visualize this model. We will use the $geom\_smooth()$ function, which adds a smoothed mean line that fits best to the data points (a regression line in our case) to the data points.

```
# Plotting the regression line for all people
ggplot(my_data, aes(x=exercise, y=body_fat)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "green") +  # line for all people
  labs(x='Exercise (hours per week)', y='Body Fat (%)',
       title='Regression Model of Exercise vs Body Fat for All People')
```

The green line is the estimated regression line, with intercept $\hat{\beta}_0 = 32.18$ and $\hat{\beta}_1 = -2.42$.

- Goal: Explore the relationship between TOEFL scores and the chance of graduate admission using a simple linear regression model.
- Instructions:
  - Download the dataset from Kaggle: https://www.kaggle.com/datasets/mohansacharya/graduate-admissions
  - In Rstudio, load the dataset and explore its contents.
  - Create a scatter plot of TOEFL Scores against Chance of Admit.
  - Build a simple linear regression model using TOEFL Scores to predict Chance of Admit.
  - Interpret the results.

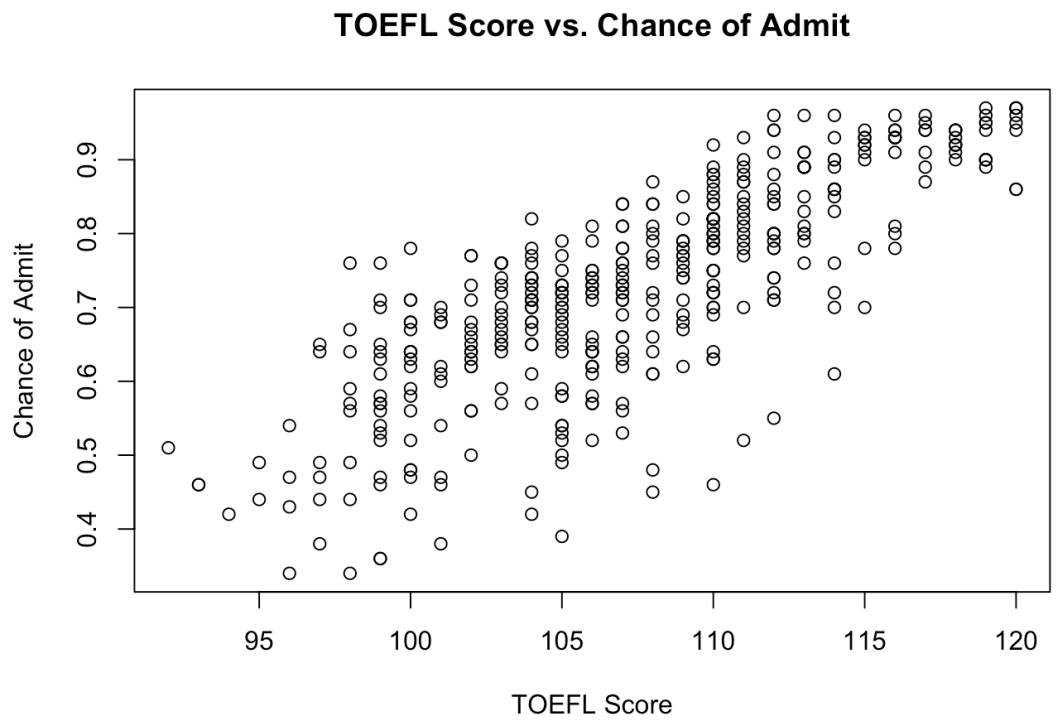
Exercise: Explore the dataset

```
# Glipmpse the dataset
glimpse(admission_data)
```

```
## Rows: 400
## Columns: 9
## $ Serial.No.        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1…
## $ GRE.Score         <int> 337, 324, 316, 322, 314, 330, 321, 308, 302, 323, 32…
## $ TOEFL.Score       <int> 118, 107, 104, 110, 103, 115, 109, 101, 102, 108, 10…
## $ University.Rating <int> 4, 4, 3, 3, 2, 5, 3, 2, 1, 3, 3, 4, 4, 3, 3, 3, 3, 3…
## $ SOP               <dbl> 4.5, 4.0, 3.0, 3.5, 2.0, 4.5, 3.0, 3.0, 2.0, 3.5, 3.…
## $ LOR               <dbl> 4.5, 4.5, 3.5, 2.5, 3.0, 3.0, 4.0, 4.0, 1.5, 3.0, 4.…
## $ CGPA              <dbl> 9.65, 8.87, 8.00, 8.67, 8.21, 9.34, 8.20, 7.90, 8.00…
## $ Research          <int> 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1…
## $ Chance.of.Admit   <dbl> 0.92, 0.76, 0.72, 0.80, 0.65, 0.90, 0.75, 0.68, 0.50…
```

Exercise: Data visualization

```
# Scatter plot of TOEFL Scores vs. Chance of Admit
plot(admission_data$TOEFL.Score, admission_data$Chance.of.Admit,
     xlab = "TOEFL Score", ylab = "Chance of Admit",
     main = "TOEFL Score vs. Chance of Admit")
```



Exercise: Linear regression model

```
# Display model summary
summary(model)$coefficients
```

```
##               Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -1.2734005 0.0774216975 -16.44759 9.443661e-47
## TOEFL.Score  0.0185993 0.0007196601  25.84456 3.634102e-87
```

- What is the relationship between TOEFL score and the chance of graduate program admission?

Exercise: Linear regression model

```
# Display model summary
summary(model)$coefficients
```

```
##               Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -1.2734005 0.0774216975 -16.44759 9.443661e-47
## TOEFL.Score  0.0185993 0.0007196601  25.84456 3.634102e-87
```

- What is the relationship between TOEFL score and the chance of graduate program admission?

The positive coefficient suggests a positive relationship: higher TOEFL scores are associated with a higher chance of admission. More specifically, for each additional unit increase in the TOEFL score, the chance of graduate program admission increases by approximately 0.0186 units.

Exercise - Linear regression model

```
# Display model summary
summary(model)$coefficients
```

```
##               Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -1.2734005 0.0774216975 -16.44759 9.443661e-47
## TOEFL.Score  0.0185993 0.0007196601  25.84456 3.634102e-87
```

- Is the relationship significant?

Exercise - Linear regression model

```
# Display model summary
summary(model)$coefficients
```

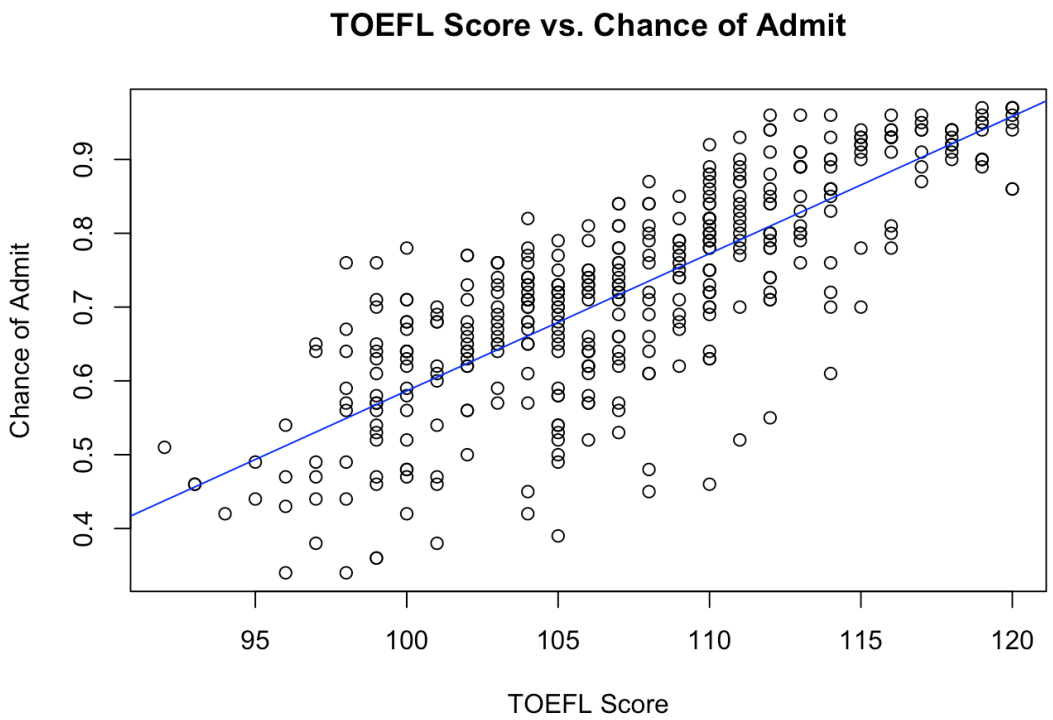
```
##               Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -1.2734005 0.0774216975 -16.44759 9.443661e-47
## TOEFL.Score  0.0185993 0.0007196601  25.84456 3.634102e-87
```

- Is the relationship significant?

Yes, the relationship between TOEFL score and the chance of graduate program admission is statistically significant. This is evident from the p-value for "TOEFL.Score", which is very close to zero (3.634102e-87), thus suggesting that the relationship is unlikely to be due to random chance.

Exercise - Model visualization

```
# Scatter plot with regression line
plot(admission_data$TOEFL.Score, admission_data$Chance.of.Admit,
     xlab = "TOEFL Score", ylab = "Chance of Admit",
     main = "TOEFL Score vs. Chance of Admit")
abline(model, col = "blue")
```



Today

Learning goals

- Understand the basic concepts of linear models.
- Learn how to interpret and analyze the output of a linear regression model.
- Understand linear model functions in R.
- Familiarize yourself with the `lm()` function in R for creating linear models.
- Gain insight into accessing and interpreting estimated coefficients using the `summary()` function.
- Apply your knowledge to a practical exercise to predict outcomes using linear regression models.